Assessing the Effect of Low-level Lead Exposure[1]

Richard A. Levine Chen-Hai Andy Tsao Yu-Ling Tseng

BU-1284-M April 1995

ABSTRACT

The effect of lead on behavior and memory has been studied at length in the psychology literature. In most of these investigations, the researchers encounter the difficult task of analyzing a repeated measures design where an experimental unit is observed at several different times. We will review the approaches used in the literature to perform such an analysis. The techniques require strong assumptions about the covariance structure of the population from which the sample is taken. Due to the dependency between some observations inherent in a repeated measures design, these procedures are quite often not applicable. Therefore, we will consider an alternative method based on mixed models. This model provides a general framework under which we can more realistically account for the potential correlation between observations. To encourage the application of this tool by scientists and statisticians we analyze a data set concerned with determining the effect of low lead levels on performance of a delayed spatial alternation task in rats.

[1]Key Words: repeated measures, mixed model, split-plot, general linear models (GLM), Proc Mixed, application

AMS 1990 Subject Classification: Primary 62J12; Secondary 62P15

Richard A. Levine is Ph.D. student, Statistics Center, Cornell University, Ithaca, NY 14853. Chen-Hai Andy Tsao is Assistant Professor, Chung Cheng University, Jia-Yi, Taiwan. Yu-Ling Tseng is Research Associate, Institute of Statistical Science, Academia Sinica, Taipei, Taiwan.

1 Introduction

We are considering the problem of the analysis of an experiment conducted for the purpose of determining the effect of low level lead dosages on memory or behavior in rats. It has been established that high doses of lead cause a variety of adverse effects on the human body. For example, lead exposure has been significantly associated with cognitive and neuropsychological functions such as attention, distractibility, intellectual and academic performance, learning, memory, and reasoning, as well as coordination, motor function, reaction time, speech and language disorders, and behavioral problems (for example see Davis et al. (1990), Dietrich et al. (1993), Hunter et al. (1985), Needleman (1994), and White et al. (1993)). However, the lower limit of hazardous lead exposure for humans is unknown.

Unfortunately, it is not a trivial matter to run an epidemiologic study to determine the effect of low levels of lead on humans. For example, ethical considerations prevent randomized experiments on human subjects and the large number of confounding factors present in a general population study creates inferential difficulties. As an alternative, scientists perform experiments on animals in a laboratory and attempt to extrapolate results to humans. These experiments allow for controlled environments which circumvent the obstacles confronted in human studies. Of course, there is much controversy over species extrapolation from animals to humans (for example, see Freedman and Zeisel (1988)). However, our concern is in the inferences drawn from the animal experiments prior to extrapolation. Discrepancies in the conclusions made from different animal studies still leaves the question of maximum allowable lead levels with negligible adverse effects unanswered. Though many problems in conducting such experiments arise from psychological considerations such as how to define and obtain evidence to indicate a deficit in memory or cognitive behavior, a major predicament is the statistical analysis of such data.

The most pronounced statistical difficulty in lead studies is that of repeated measures. In general, the term repeated measures defines a design in which an experimental unit is observed under each level of a particular factor or treatment.

For example, an experimenter may test a single subject from a lead treatment group on a memory task over a number of days. Here the observations over the time factor are the repeated measurements.

The advantages of a repeated measurements design are, firstly, we can determine the effect of a treatment over another factor such as time. Time trends are better established by observing an individual over time rather than comparing observations on different individuals at specified points in time (Neter, Wasserman, and Kutner (1990) chapter 28, Snedecor and Cochran (1989) section 16.6). Secondly, the subjects essentially serve as their own control with respect to the factor over which the repeated measures are made. In other words, by observing a single subject over a series of days, say, effects due to time are determined by comparing observations for each individual. Hence, only within subject variation and not between subject variability is included in the experimental error resulting in more precise time effect estimates (Koch et. al (1988)). Thirdly, the study needs only a small number of subjects for implementation. Consequently, the repeated measures design may be less costly and perhaps easier to utilize when few experimental units are available (Neter, Wasserman, and Kutner (1990)).

However, the repeated measures design has a number of disadvantages which create statistical problems from the analysis perspective. The most serious of these drawbacks is that the repeated observations over a factor are usually not independent, a common supposition made in data analyses. For example, measurements on an individual over time may be correlated in such a way that responses closer together in time are more similar than those further apart in time. Furthermore, the repeated measures factor may be confounded with other experimental components implying the effects due to each individual factor are inseparable. For example, if the response being measured in a study is based on some mental task, it is hard to distinguish time effects from learning or memory effects.

In this paper we will consider statistical techniques for overcoming the problems inherent in data from a repeated measures design. We are particularly interested

in studying the mixed model, a flexible general linear model as presented in Laird and Ware (1982). Throughout, we have in mind an application whereby data is collected on rats over time for determining the effect of low lead levels on memory or behavior. In Section 2, we study several analysis tools used in the literature dealing with the effects of lead on memory (which we call the lead effects literature from here onward) via repeated measures designs and describe the advantages and disadvantages of each technique. In Section 3, we introduce an alternative statistical method based on mixed models which better takes into account the correlation between observations within a subject. Section 4 then analyzes data utilizing the mixed model methodology and shows how recent computer software makes the implementation of the procedure manageable. In Section 5, we discuss areas for future research and summarize the ideas presented in the paper.

2 Review of Methods

Consider a study aimed at determining the effects of low lead dosages on memory or behavior in animals. In particular, imagine we have assigned subjects to one of two treatment groups- low lead and zero lead (a precise experiment will be described later). Each subject is measured for performance at a particular task over a number of days. The hope is that comparisons between the two treatment groups of subjects' performance on the task will establish if any behavior abnormalities arise from exposure to low lead dosages. Under this experimental design we are confronted with a repeated measures problem whereby the same individual is observed at each level of the time factor (i.e., multiple observations on the same subject over time). The repeated measures design is common in the literature focusing on the effects of lead on behavior and memory. Upon a search of these works, we have discovered the utilization of a number of different statistical techniques for analyzing data from this type of study design. In each procedure, the correlation between observations within the same subject is addressed by somehow summarizing the responses over time or avoiding the issue altogether. These approaches, as we will see

in Section 2.1, simplify the analyses but lose some information available in the data if not invalidating the methodology entirely.

There are more appropriate methods for analyzing data from a repeated measures design which account for the correlation between observations more directly and hence produce better inferences on the various effects of interest than those described in Section 2.1. An overview of these alternative procedures is given in Section 2.2. In particular, we consider the general linear model , and more specifically the mixed model, as described by Laird and Ware (1982) and Ware (1985). This model subsumes most of the analysis approaches we discuss in this review section and is flexible in terms of modeling covariance structure.

Our main goal in this paper is to encourage the utilization of the mixed model in practice. To this end, we attempt to show the advantages of thinking about statistical analyses of repeated measures data through the mixed model framework. Such a mind set allows for flexibility in terms of model choice. Moreover, though not all mixed models are feasible computationally, many can be easily implemented through new statistical software packages. Our presentation emphasizes these points by first, in Section 2.1, reviewing the commonly applied statistical techniques in the lead effects literature. Then, in Section 2.2, we provide an informal motivation for extensions of these tools to more biologically practical general linear and mixed models of Laird and Ware (1982). In Section 3 we detail the theory behind the mixed model and describe how such a model may be fit. We additionally detail how some of the approaches discussed in Section 2 may be written as a mixed model. Finally, in Section 4, we illustrate how to apply new computing programs for fitting and interpreting a mixed model to determine whether there exists a significant effect of low lead levels on behavior or memory in rats.

2.1 Standard Approaches in the Lead Effects Literature

One commonly used approach for analyzing data from a repeated measures design in the lead effects literature is to ignore the repeated measures problem completely.

We may try fitting a two-way analysis of variance (ANOVA) model with time and treatment as the explanatory variables. Under the standard ANOVA assumptions of independent errors sampled from a normal distribution with constant variance across observations, the effects of lead treatment and time on performance of the specific task can be determined. This model is simple to implement and easy to interpret. However, the observations on a single individual over time will most likely be correlated. It seems unreasonable to assume that the performance of a subject on a task on one day will be independent from his/her performance on a subsequent day. Hence, the test statistics for the treatment and time effects do not have F distributions and thus the test results and conclusions drawn from the fitted model will not be valid.

As an alternative to the two-way ANOVA model, if we can assume the effect of time on the response is not significant, then we can summarize the data over time, say, by averaging the responses over each individual. Consequently, we replace the repeated measures with one summary measure per subject and hence can perform a valid one-way ANOVA with treatment as the explanatory variable (assuming observations between individuals are independent- a reasonable supposition). Of course, this simplification results in a loss of information pertaining to time if the presumption of no time effect is false. In other words, by averaging the response for a single individual over the days of the experiment, we can not study nor take account of time trends in our analysis. On the other hand, this "averaging- ANOVA" model is the easiest to implement and may be a good first step at exploring and analyzing the data prior to attempting one of the more complicated techniques described later. For applications in the lead effects literature of the averaging-ANOVA idea see [1], [2], [7], [16], [24], and [43].

Simplifying the repeated measures over time by means of a summary statistic such as an average is not the only technique for overcoming the correlation between observations within a subject. Another approach commonly utilized in the lead effects literature considers each day as a separate experiment. Hence, if the experi-

ment is run over T days, we can fit T one-way ANOVA models on the observations for a specific day and test for treatment effects. This procedure requires the standard ANOVA assumptions for data collected on a particular day; i.e., independent observations sampled from a normal population with constant variance.

The individual ANOVA models are valid in that the ANOVA assumptions should hold on a particular day (i.e., the supposition of independent observations on different subjects seems reasonable) and we can perform T legitimate tests for treatment effects. Furthermore, it is easy to calculate the T ANOVAs. However, there are a number of disadvantages to the "many-ANOVAs" methodology. First of all, since we are analyzing each day separately and the observations from the same individual are correlated over time, we can not control for the time effect in our analysis nor study time trends by any standard statistical technique.

Second, we need to consider issues such as performing multiple tests and drawing an overall conclusion from the group of tests. For example, we can adjust the α-level for each individual test using Bonferroni corrections (Neter, Wasserman, and Kutner (1990) section 5.1) to guard against a large Type I error for the family of tests. But after adjusting for the large number of treatment comparisons being performed, it is not clear how the results can be combined across tests that are evaluated on the same experimental unit.

Third, we can not directly model potential interactions between time and treatment. This interaction might provide more insight into the effect of the lead treatment on memory or behavior than the treatment effects alone. For example, the interaction term may indicate that the difference in performance between the lead and control groups is increasing over time. Such a nonparallelism in the treatment group responses can be interpreted as a behavioral effect (attitude change later in the study for lead group) or increase in memory impairment over experimental days. Since we have a separate ANOVA for each day, we can compare the parameters to study time by treatment interactions. But such a procedure is not as straightforward as the ideas discussed later. As with the averaging-ANOVA approach, though,

the many-ANOVAs technique is easy to implement and hence may be a good initial attempt at understanding the data. The many-ANOVAs idea is used by [2], [4], [5], [7], [36], [38], [50], and [54] in the lead effects literature.

The previous two methods solve the repeated measures problem at the cost of ignoring the effect of time on the response. We would like a procedure which simplifies the problem by perhaps summarizing the repeated measures over time while evaluating or taking advantage of the time trends. One such approach sometimes used in the lead effects literature is to summarize the repeated measures by way of a parametric curve. For example, we can fit a line through the data obtained from a subject with the task performance (response) as the dependent variable and time as the explanatory variable,

$$\text{Performance} = \alpha + \beta \cdot \text{Time}.$$

The estimated coefficient of the linear component, $\hat{\beta}$, can then be substituted for the response variable for that particular subject. As with the averaging-ANOVA approach, by representing the observations over time with one summary measure, an ANOVA can be performed to test for treatment effects. More generally, for each subject we can summarize the profile over time with a parametric curve. The estimated parameters can then be used as dependent variables in a multivariate analysis of variance (MANOVA) on treatment or as separate ANOVAs on treatment for each parameter.

This summary technique essentially divides the problem into two parts: one analyzing the time effect and one studying the treatment effect. In phase one, a parametric curve for each subject's responses over time is assumed. In phase two the standard ANOVA assumptions are needed. Hence, in a sense, we are first summarizing the data over time (i.e., eliminate the time trend) and then study the treatment effect in a separate analysis.

There are many advantages to summarizing the repeated measures by means of a parametric curve. First of all, by replacing the repeated measures with one summary measure per subject we do not need to worry about the correlations between

observations within individuals. Therefore the ANOVA tests in phase two will be valid assuming observations on different individuals are independent; a reasonable supposition. Secondly, the choice of a parametric curve in phase one calls for expert knowledge about the situation under consideration. Therefore, unlike the averaging-ANOVA approach where averages seem to be used as a summary technique merely for simplicity, the motivation behind summarizing the time trend by a fitted curve is more intuitive and scientifically based. Furthermore, phase one allows us to measure and analyze a time trend in the responses. As with the many-ANOVAs approach, we can study treatment by time interactions through the parameter β if it depends on treatment.

Unfortunately, this methodology relies heavily on the class of curves upon which we decide prior to the analysis. Not only are we restricted to study time trends within only the chosen class, but if our selection is incorrect the conclusions we draw from the fitted model may be misleading or invalid. Of course, we might be able to perform some statistical tests to establish the validity of our model; but we will again face the difficulty of dealing with the correlation between observations on an individual. Since the main goal of this summary technique is to provide a simple means for avoiding the impediments inherent in repeated measures, rather than complicate the issue further, we would prefer to keep the curve selection process subjective using expert opinions. If the statistician is truly unhappy with such an approach, instead of solving the problem through model validation in this framework, the mixed model methodology discussed later is probably a more tenable approach. In addition to the problems of selecting a parametric curve, by studying the time trends separately from the treatment effects, we may lose relevant information in the data that would be better represented by a combined model.

Under the assumption that our expert opinion is correct and the presumed functional form of the time trend is accurate, overall this method lends itself to an easy implementation and provides for straightforward interpretations of the data analysis from a repeated measures design. A number of papers utilize this technique

for analyzing lead data. For example see [11] and [12] (use linear term of cubic polynomial); [3], [10], [9], [20], [36], [46], [50], [47], [48], [49], [51], [52], [53], and [59] (use intercept- essentially averaging).

For another commonly employed approach which summarizes the data across the repeated measures but allows for a test of time effects and treatment by time interactions, recall that we have applied a lead treatment to each subject (experimental unit) and observations are then taken on the subjects over time. Hence, this study can be thought of as a split-unit design where the subjects are "whole-plots," to which one treatment is applied, and each subject is split into units where the effects of a second factor are observed (Mead (1988), Neter, Wasserman, and Kutner (1990) chapter 28, Snedecor and Cochran (1989) section 16.16). Here, the split-unit or within subjects variable is time. The split-unit design assumes observations on different subjects are independent and, prior to selecting the observations within the experimental units are correlated; reasonable assumptions in our situation. Therefore we can conduct the analysis of our data from the repeated measures design by way of a split-unit type of analysis. Subsequent to choosing subjects for the experiment, the individual observations are assumed independent with errors distributed normally with constant variance for a standard split-unit analysis to be done.

Notice that the split-unit approach to the repeated measures problem is just another way of summarizing the data within a subject over time. At the whole-plot or subject level, we are comparing average treatment effects over time as in the averaging-ANOVA approach. Therefore any tests on treatment effects are valid since we remove the correlation across repeated measures on each subject. Unlike the averaging-ANOVA idea, though, the split-unit analysis allows for a test of time effects at the split-unit level, if the underlying design assumptions above are correct. Hence, we can measure trends in the response over time and treatment by time interactions. Furthermore, the analysis is not hard to implement. In particular, if there are only two levels of time (e.g., the experiment is run on two days), the split-unit F tests are paired t-tests (correlation does not matter when comparing

two points).

Unfortunately, the correlation structure of the observations required by a split-unit type analysis may not hold for data collected from a repeated measures design. It can be shown that in the standard model for analyzing data from a split-unit design, any two observations from a subject are assumed to have the same correlation (Cochran and Cox (1957) section 7.12, Neter, Wasserman, and Kutner (1990) chapter 28). However, equicorrelation is probably not a reasonable assumption for within-subject observations. For example, we might expect a subject's response to be more highly correlated at points closer in time than points far apart in time (due to temporal changes in attitudes say). If the equicorrelation supposition does not hold, the F-tests on the split-unit level are invalid. An additional problem is that, in general, tests on the whole plot level are not as precise as tests on the split-unit level (Cochran and Cox (1957) section 7.12, Lentner and Bishop (1993) section 11.2). In our situation, the test for treatment effects is most important, but treatment is the whole-plot factor. Therefore, an alternative analysis with a more powerful test for treatment effects may be desired.

These two disadvantages are not devastating in that the distribution of the test statistic for the time effect can be approximated by an F distribution using Greenhouse and Geisser (1959) techniques, if equicorrelation does not hold. Furthermore, the less precise test of treatment effects err on the conservative side (i.e., larger Type II errors for whole-plot tests than split-unit tests). Hence, if we find a significant treatment effect, we can be satisfied. Nevertheless, considering that the method of summarizing the data by averaging over time within a subject creates potential difficulties and it does not take advantage of expert knowledge of possible time trends, the curve-fitting procedure seems to be the preferred technique over the split-unit analysis. Overall, due to these two issues and the additional fact that the averaging-ANOVA and many ANOVAs procedures can not control for or measure time effects, the equally easy to implement curve-fitting idea seems best among the four methods of analysis presented. Split-unit methodologies are used by [5], [6],

[16], [33], [34], [35], [36], and [38] in the lead effects literature.

All the statistical methodologies discussed in this section address the correlation problem between observations within a subject by either summarizing the data over the repeated measure or somehow ignoring the dilemma altogether. Each leads to an algorithm which is easy to implement in any standard statistical package. Yet, they all require restrictions on the correlation between observations that are either not met by data collected from a repeated measures design or do not take full advantage of information on a subject over time. Therefore, a statistical technique that directly models the correlation structure encountered in repeated measures data may allow us to utilize the extra information we obtain on an individual to analyze time trends and perform more precise tests of treatment effects than otherwise possible. In the next subsection we consider a few alternative approaches which may better model the underlying correlations in longitudinal data.

2.2 Alternative Approaches

The techniques discussed in Section 2.1 can be categorized as either growth curve models or ANOVA models. In this subsection we will study extensions of the models from these two categories as a further attempt to account for the inherent correlation between repeated measurements on the same individual. We will first consider how growth curve analyses can be modified to provide for a more realistic model. Then we will outline a unified approach based on general linear models which subsumes the previously discussed methods as special cases. Details behind the formulation of these models are described in subsequent sections.

A growth curve provides a means of representing the relationship between a response of interest and the repeated measures factor on a particular individual. For example, in Section 2.1, we discuss fitting some parametric curve or polynomial describing the performance of a subject on a task over time. As introduced by Wishart (1938), we would reduce the polynomial growth curve into a set of estimated coefficients and compare these coefficients across groups of individuals to determine,

say, treatment effects.

Unfortunately, a polynomial growth curve, though simple to fit mathematically, may not be realistic from a biological viewpoint (Sandland and McGilchrist (1979)). We need a flexible model based on biological as well as statistical considerations. Models built on differential equations may be able to better approximate complex real life phenomena and hence allow for meaningful interpretations of parameters. In particular, these models consider changes in the response continuously over time. Since many biologic processes are studied in continuous rather than discrete time, the differential equation models are then more realistic (Dwyer et al. (1991) chapter 3). Furthermore, the differential equations can be formed into a wide range of shapes and include terms to model the covariates relevant to the response of interest. The models can also incorporate a feedback term which allows measurements at particular time points to depend on previously observed outcomes (see Dwyer et al. (1991) chapter 3 for details). Therefore, as compared to polynomial growth curves, the differential equation models provide a larger class of functions from which we may choose for a more appropriate fit of our data and the underlying biological processes. Once fitting the differential equation for modeling of the time effect (i.e., the effect over which the repeated measures are taken), we can compare different treatment groups by way of a multivariate analysis of variance (MANOVA) on the estimated parameters for the individual growth curves (Sandland and McGilchrist (1979)).

As with any parametric growth curve, however, the differential equation approach requires the researcher to define the experimental dynamics through some functional relationship. Though potentially any function may be chosen, the scientist must limit the complexity so as to allow for feasible estimation and interpretation of the model parameters. We can eliminate the dependence on functional form by performing nonparametric curve estimation. This procedure is more flexible in that the only constraints on the function are in terms of smoothness and differentiability. We do not need to make assumptions as to the actual shape of the function. As

with the other growth curve methods, upon fitting a nonparametric curve, we obtain some characteristic of the curve or its derivatives and compare these across the various treatment groups through a multivariate analysis technique. See Müller (1988) for details on applying nonparametric models to longitudinal data.

Note that in each of these growth curve methods, we have suggested summarizing the curves through a vector of parameters representing characteristics of the function. Then, we compare curves, say between different treatment groups, by performing a multivariate analysis (MANOVA, cluster analysis, or discriminant analysis to distinguish significantly different groups). However, we can carry the nonparametric idea a step further. Again, the classical approaches require distributional assumptions (typically multivariate normal) on the response variable. If we perform a nonparametric test, these restrictive suppositions are not required. Koziol et al. (1981), Wei and Johnson (1985), and Davis and Wei (1988) present procedures for nonparametric comparisons of growth curves.

Let us now consider placing all the approaches discussed in this review section under one "umbrella" model. Such a generalization would give the practitioner a common framework under which to choose and fit a model for a given situation. The unified approach we will study is the general linear model (GLM) as presented in Laird and Ware (1982). For now we will give an informal overview of the methodology and leave the details for Section 3. Let us motivate the GLM idea through scrutiny of the growth curve analysis. Notice in the case of fitting a polynomial curve to serial measurements that we write the mean response for individual i as a function of additive terms each increasing in a power of t,

$$\mathrm{E}(y_i) = a_{0i} + a_{1i}t + a_{2i}t^2 + \ldots + a_{pi}t^p.$$

If we want to compare different growth curves, we can test whether the a_{ji} are equal across all individuals, $i = 1, \ldots, n$, for each coefficient $j = 1, \ldots, p$. Hence we are in a general linear model framework where the response is a linear function in the coefficients of the polynomial in time. The GLM characterization of growth curves was suggested and discussed by Potthoff and Roy (1964) and Rao (1965). In

this setting we can also include covariates, both time dependent and independent (Grizzle and Allen (1969)) and allow for different curves over treatment groups say.

Of course, we are not limited to representing the effect of time on the response as a polynomial. More specifically, the GLM can account for and model both between and within subject variation as long as the expected response is a linear function of the parameters of interest. Additionally, it can also handle analyses of data from unbalanced designs, measurements taken repeatedly over unequally spaced intervals (e.g. some individual observed every day, others every other day, etc.), and data sets with missing observations (Ware (1985)). Again, we will discuss this matter more fully in Section 3. Furthermore, we will formally detail how the previous methods can be written as GLMs and, in Section 4, illustrate choosing a mixed model and fitting it to data dealing with the effects of lead on behavior. For another application of the mixed model in the lead effects literature see Waternaux, Laird, and Ware (1989) and Dwyer et al. (1991) chapter 4 (also written by Waternaux and Ware on the same analysis).

One drawback of the GLM, as we will discover later, is that we need to assume a special form for the covariance structure in the responses. We can generalize further, though, by assuming an unstructured multivariate model for the response. However, such generalizations can become mathematically intractable or too computationally intensive for a large covariance matrix. Additionally, it may be difficult to define individual random characteristics for the subjects with the full multivariate model (Laird and Ware (1982)).

The GLM framework can also be used to address other issues that may arise when modeling longitudinal data. First, we may wish to include a feedback mechanism, i.e., allow measurements at time t to depend on previous measurements $1, \ldots, t-1$. A general linear model similar to the Laird and Ware (1982) model but conditional on previous responses can be constructed for such a task (Dwyer et al. (1991), chapter 5). Secondly, many situations are fit better by functions nonlinear in the parameters. Nonlinear models are more complicated but can be fit in

certain circumstances. See, for instance, Jones (1993) chapter 7. Thirdly, we have been implicitly assuming a continuous response in the various approaches discussed. However, the GLM can be extended to discrete data by choosing an appropriate "link" function between the mean response and the explanatory variables. In particular see Fitzmaurice et al. (1993), Dwyer et al. (1991) chapters 7-9, Stanek and Diehl (1988), and Zeger et al. (1988) to name a few.

In the remainder of the paper we will discuss the mixed model methodology as an application of the general linear model to a repeated measures situation. Most models for serial measurements can be considered from a GLM framework (Laird and Ware (1982)). Therefore, by exploring the mixed model, we hope the reader can extend and apply it to many different experimental situations arising in practice. Furthermore, due to new computer packages, the mixed model is not hard to implement and provides a more thorough method for answering the question of the effect of lead on behavior in animals.

3 Mixed Model Theory

We are studying the relationship between certain lead dosages and memory in rats where repeat observations are taken on an individual over time. Therefore, we would probably model the effects of the lead treatment as fixed since interest lies only in those levels of lead being applied during the experiment. However, in modeling changes in the response over time we are not necessarily interested in the effects due to those specific days on which the experiment is run but a population of time effects on memory. Hence, we may wish to consider the time effects as random rather than fixed (for further discussion of this idea see Section 4). Additionally, we would like to better model the correlation structure between observations on the same individual than accomplished by the techniques discussed in Section 2. The mixed model methodology allows us to incorporate these factors into the analysis of data from a repeated measures design. In this section we will review the theory for implementing the mixed model and then, in Section 4, we will apply it to the lead

problem.

The general mixed model can be written (Searle, Casella, and McCulloch (1992) section 4.6)

$$\mathbf{Y}|\mathrm{u} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathrm{u} + \boldsymbol{\epsilon} \tag{1}$$

$$\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}\ ,\ \mathrm{cov}(\boldsymbol{\epsilon}) = \mathbf{R}$$

and

$$\mathrm{E}(\mathrm{u}) = \mathbf{0}\ ,\ \mathrm{cov}(\mathrm{u}) = \mathbf{G} \tag{2}$$

where, for N data points, f fixed effects, and r random effects, the $N \times 1$ vector $\mathbf{Y}$ is the response of interest (e.g., rat memory), the $f \times 1$ vector $\boldsymbol{\beta}$ represents the fixed effects (e.g., lead treatment), and the $r \times 1$ vector u represents the random effects (e.g., time, time by treatment interaction). Furthermore, $\mathbf{X}$ and $\mathbf{Z}$ are known design matrices of dimensions $N \times f$ and $N \times r$ respectively, $\boldsymbol{\epsilon}$ is an $N \times 1$ random vector (e.g., sampling error), $\mathbf{R}$ is an unknown $N \times N$ covariance matrix corresponding to the random variable $\boldsymbol{\epsilon}$, and $\mathbf{G}$ is an unknown $r \times r$ covariance matrices corresponding to the random variable u.

The mixed model has a number of interesting features. Notice that it is divided into two parts, (1) and (2). The first stage models the response conditional on the realization of the random effect u. In other words, at the point of data collection, we have a sample of factor levels representing the population of treatments corresponding to the random factor. The effects of these levels on the response are then fixed, though unobservable. The matrix $\mathbf{R}$ takes account of some of the correlation between observations, depending on the structural assumptions we may place on it. Overall, at the first stage (1), we are assuming a linear and additive relation between the response $\mathbf{Y}$ and the fixed effects and the realization of the random effects, $\boldsymbol{\beta}$ and u (i.e., a general linear model). Furthermore, the random error term is assumed to have mean zero with some covariance matrix $\mathbf{R}$. The zero mean supposition is typical for ANOVA-type designs. However, we are not constrained to the usual ANOVA assumption of independent error terms where $\mathbf{R} = \sigma^2\mathbf{I}$ with $\mathbf{I}$ being the identity matrix and σ^2 an unknown variance. Typically though, $\mathbf{R}$ is block diagonal since the

observations within a subject are dependent and observations between subjects are independent. Furthermore, as in general linear models, ϵ is often presumed to have a multivariate normal distribution (Searle, Casella and McCulloch (1992)).

The second stage of the model, (2), places distributional assumptions on the random variable u. We assume u has a mean of zero, a common supposition for random effects. Additionally, in this model, we study the correlation between random effects through the matrix G. In summary, the mixed model assumes the random effect u (e.g., time) is sampled from some distribution with zero mean and covariance structure G. Subsequent to choosing the levels of the random factor (i.e., given the realization of the random effect), the responses are distributed normally with a mean represented by some linear and additive function of the effects u and β with correlation structure R.

Some of the analyses discussed in Section 2 can be represented by the mixed model (1) and (2). For example, the averaging-ANOVA and many ANOVAs methodologies can be written in terms of the mixed model if we eliminate u (since these techniques ignore the time effect) and assume $\mathbf{R} = \sigma^2\mathbf{I}$ (ANOVA independence assumption). Y then signifies the average response for each subject in the averaging-ANOVA case and the responses for all individuals on a single time unit in the many ANOVAs scenario. The split-unit design can be described by the mixed model as follows. The rat is the whole-plot experimental unit split by the time factor. The assumptions are different than those presented before in that the effect of rat, not time, is assumed random. Let $\beta = [\tau', \mathrm{t}', (\tau \times \mathrm{t})']'$ be the vector of fixed effects where τ is the vector of lead treatment effects, t is the vector of time effects, $\tau \times \mathrm{t}$ is the vector of interactions between treatment and time, and $'$ signifies the transpose of a vector. Furthermore, define u as the vector of effects corresponding to the r rats. Then, if $\mathbf{R} = \sigma^2\mathbf{I}$ and $\mathbf{G} = \sigma_R^2\mathbf{I}_r$, where σ_R^2 is the variance component of the rat variable and $\mathbf{I}_n$ is an $n \times n$ identity matrix, (1) and (2) will fit data from a split-unit design.

An analysis based on growth curves can also be written as a mixed model. For

example, assume we have n rats divided into three lead treatment groups and measured q times each. Further assume the variation in the response for an individual subject over time is represented by a $p-1$ degree polynomial and that this polynomial is the same for each individual in a particular lead treatment group. Following the work of Vonesh and Carter (1987), we can write this growth curve model in terms of (1) and (2) by letting

$$\mathbf{Y}_i = \mathbf{A}_i \left(\mathbf{I}_p \otimes \mathbf{V}_i\right) \mathrm{u}_i + \mathbf{A}_i \left(\mathbf{I}_p \otimes \mathbf{V}_i\right) \boldsymbol{\lambda} + \boldsymbol{\epsilon}_i \tag{3}$$

where $\mathbf{Y}_i$ is a $q \times 1$ vector containing the responses for the ith individual at each of the q time points,

$$\mathbf{A}_i = \begin{pmatrix} 1 & t_{i1} & t_{i1}^2 & \dots & t_{i1}^{p-1} \\ & & \vdots & & \\ 1 & t_{iq} & t_{iq}^2 & \dots & t_{iq}^{p-1} \end{pmatrix}$$

is a $q \times p$ matrix containing the various powers of time for the polynomial at the q different measurement points, $\mathbf{V}_i = (\mathbf{I}(\mathrm{TRT}_i = 1), \mathbf{I}(\mathrm{TRT}_i = 2), \mathbf{I}(\mathrm{TRT}_i = 3))$ is a 1×3 vector where $\mathbf{I}(\mathrm{TRT}_i = j)$ is the indicator function that the ith individual is in the jth treatment group, and $\boldsymbol{\lambda} = (\lambda_{10}, \lambda_{20}, \lambda_{30}, \lambda_{11}, \lambda_{21}, \lambda_{31}, \dots, \lambda_{1p}, \lambda_{2p}, \lambda_{3p})^T$ is a $3p \times 1$ vector where λ_{jk} is the coefficient for t^k for the jth treatment group growth curve (polynomial). Furthermore, $\boldsymbol{\epsilon}_i \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I}_q)$ so $\mathbf{R}_i = \sigma^2 \mathbf{I_q}$ and $\mathrm{u}_i \sim \mathrm{N}(\mathbf{0}, \mathbf{D})$ is a $3p \times 1$ vector of random effects. The symbol $\otimes$ denotes the direct product of the two matrices $\mathbf{I}_p$ and $\mathbf{V}_i$. In other words, for the jth time point of measurements on rat i who is in treatment group one, $\mathbf{V}_i = (1, 0, 0)$ and

$$E(y_{ij}) = \lambda_{10} + \lambda_{11} t_{ij} + \lambda_{12} t_{ij}^2 + \ldots + \lambda_{1p} t_{ij}^{p-1}.$$

Therefore, we can compare the coefficients λ_{jk}, $k = 1, \dots, p$, across the three groups $j = 1, 2, 3$ to determine if the response significantly differs between the three lead treatments.

The linear model (3) can be applied to even more general situations. For example, we are not restricted to an experiment whereby all individuals are observed at the same times or on the same number of occasions. We can vary $\mathbf{A}_i$ to account for

such unbalanced designs. Additionally, we can include time independent covariates such as initial body weight by using $x_i\mathbf{V}_i$ instead of $\mathbf{V}_i$ where x_i is the body weight of the ith individual at the start of the study. Furthermore, we can consider time dependent covariates such as body weights on each of the q days of the experiment by replacing $\mathbf{A}_i$ with $\mathbf{B}_i\mathbf{A}_i$. Here $\mathbf{B}_i$ is a $q \times q$ matrix whose rows either equal $\mathbf{x}_i = (x_{i1}, \ldots, x_{iq})$ such that x_{ij} is the body weight of the ith rat at the jth time point or have entries all equal to one. Therefore, if every row of $\mathbf{B}_i$ is equal to $\mathbf{x}_i$, then this formulation will represent the average response as

$$E(y_{ij}) = x_{ij} \cdot (\lambda_{10} + \lambda_{11}t_{ij} + \lambda_{12}t_{ij}^2 + \ldots + \lambda_{1p}t_{ij}^p).$$

Note, too, by setting all entries in a specific row of $\mathbf{B}_i$ to one we can eliminate undesired interactions between the covariates and time. These general modeling schemes allow for flexibility not offered by the standard growth curve models (Ware (1985)).

The mixed model, then, is a generalization of some of the models discussed in Section 2. The covariance matrix $\mathbf{R}$ provides a means for modeling any type of correlation due to repeated measurements taken on the same subject. Additionally, we can model the covariance structure of the random effect, u, through the matrix G. However, there appears to be some disadvantages to the mixed model. Firstly, interpretation of the various model parameters may be veiled by the complexity of the model. Hopefully, though, the general linear model framework utilized will allow for similar interpretations of the model components as those made in the standard ANOVA-type models and avoid difficulties due to complexity. Secondly, the statistical analysis may be computationally intensive. With the help of software packages such as GAUSS, MATLAB, SAS-Proc Mixed, inferences about the effects and covariance elements as well as tests of fit will be accessible numerically and thus feasible.

The estimation of variance components and effects in the mixed model revolves around likelihood function methodologies. To implement these procedures we need to assume the random factors, u, have some underlying probability distribution. A

common choice is the normal distribution since the likelihood functions can then be derived or at least be computable. Two popular estimation procedures based on the likelihood function are the maximum likelihood estimator (MLE) and restricted (or residual) maximum likelihood estimator (REML). The MLE consists of maximizing the likelihood function $L(\boldsymbol{\beta}, \mathbf{R}, \mathbf{G} \mid \mathbf{Y})$ for the fixed effects, $\boldsymbol{\beta}$, and the covariance matrices, $\mathbf{R}$ and $\mathbf{G}$. Iterative methods such as the EM algorithm or Newton-Raphson (see Tanner (1993)) procedure are usually involved to calculate the estimators. Unfortunately, as pointed out in Searle, Casella, and McCulloch (1992), Section 6.6, MLEs have the undesirable property that when estimating variance components they do not take into account the degrees of freedom involved in estimating the fixed effects. As a simple example, suppose we have data $x_1, \ldots, x_n$ sampled from a normal distribution with unknown mean μ and unknown variance σ^2. The MLE of σ^2 is found to be $\hat{\sigma}^2 = \sum(x_i - \bar{x})^2/n$. The uniformly minimum variance unbiased estimator (UMVUE) is $\sum(x_i - \bar{x})^2/(n-1)$. We can think of the $n-1$ in the denominator of the UMVUE as n degrees of freedom for the n data points sampled minus one for estimating μ. The MLE ignores, in a sense, the estimation of μ by using n in the denominator when estimating σ^2. Consequently, variance components may be underestimated by MLEs. Since tests of the fixed effects are evaluated through comparisons with these estimated variances, the MLEs may lead to an exaggeration of the fixed effects (i.e., potential incorrectly deflated p-values).

The calculation of REML estimators overcomes the degrees of freedom difficulty encountered by MLEs. The REML estimator maximizes the marginal likelihood $L(\mathbf{R}, \mathbf{G} \mid \mathbf{Y})$ rather than $L(\boldsymbol{\beta}, \mathbf{R}, \mathbf{G} \mid \mathbf{Y})$, as done by the MLE. The hope is that by "integrating" out $\boldsymbol{\beta}$ to obtain $L(\mathbf{R}, \mathbf{G} \mid \mathbf{Y})$, the REML estimator will maximize a "part of the likelihood invariant to fixed effects" (Thompson (1962)). Hence, the degrees of freedom lost when estimating $\boldsymbol{\beta}$ will not be an issue in REML estimation of $\mathbf{R}$ and $\mathbf{G}$. Unfortunately, as a consequence, the REML procedure does not provide estimators for the fixed effects, $\boldsymbol{\beta}$. To obtain estimates of $\boldsymbol{\beta}$, it is common practice to use an empirical Bayes type approach and maximize $L(\boldsymbol{\beta}, \hat{\mathbf{R}}, \hat{\mathbf{G}} \mid \mathbf{Y})$ where $\hat{\mathbf{R}}$

and $\hat{\mathbf{G}}$ are the REML estimators of $\mathbf{R}$ and $\mathbf{G}$ derived from $L(\mathbf{R}, \mathbf{G} \mid \mathbf{Y})$.

Calculations of both the ML and REML estimators may be computationally intensive. Therefore, we may wish to place restrictions on $\mathbf{R}$ and $\mathbf{G}$ to make the maximizations more mathematically tractable. However, they are both based on the maximum likelihood principle which is known to have good statistical properties, particularly in asymptotic theory (see Lehmann (1983) chapter 6). With faster and bigger computers being developed, these useful statistical features outweigh the potential impracticality of employing likelihood based estimators over others. Furthermore, as we will show in the next section, software packages exist for computing likelihood estimators of β, $\mathbf{R}$, and $\mathbf{G}$ under certain mixed models of the form (1) and (2).

4 Data Analysis

To illustrate the implementation of the mixed model theory discussed in Section 3, we will analyze data from a study on the effects of chronic low lead exposure on cognitive functioning in rats. In the experiment, forty five rats are randomly assigned to three lead treatment groups consisting of 300 parts per million (ppm) sodium acetate water (no lead: control group), 75 ppm lead acetate water (low lead group), and 300 ppm lead acetate water (high lead group). The lead is administered via the rats' drinking water throughout the experimental period.

Each rat is tested on a delayed alternation task aimed at obtaining a measure of memory. The object of this task is to alternate nose-poke responses between two adjacent funnels. The animal is rewarded with a food pellet for making a response in the funnel opposite from the previous response. Proficiency on this task requires the ability to hold information (i.e., which funnel selected in the preceding nose-poke) in memory across a temporal gap between trials. An intertrial delay (0, 10, 20, or 40 seconds) is randomly imposed between trials. Prior to this testing procedure, the rats are taught the rules of the game and familiarized with the experimental apparatus. Hence we assume the responses are a measure of memory

and not learning with respect to this task.

The experimental design consists of every rat participating in one daily session, six days a week for a total of eleven sessions. Each session contains 100 trials or lasts 90 minutes, whichever comes first. The trials are run in one of ten experiment boxes to which rats are randomly and permanently assigned at the beginning of the study. Performance measurements are defined as the percent of correct responses across trials of the same delay for each session. Thus the repeated measures design is apparent in that observations are taken on the same rat over different sessions and different intertrial delays.

The mixed model from Section 3 provides a means to account for the repeated measurements within a rat and study the effects of lead on memory in this data set. For each individual observation we can consider breaking down the responses into additive components by

$$y_{ijkl} = \mu + \tau_i + d_j + s_k + \epsilon_{ijkl} \tag{4}$$

where μ is an overall mean; τ_i is the lead treatment effect for the ith group, $i \in \{1, 2, 3\}$; d_j is the delay effect, $j \in \{1, \ldots, 4\}$; s_k is the session effect, $k \in \{1, \ldots, 11\}$; ϵ_{ijkl} represents the random error component, $l \in \{1, \ldots, 45\}$; and y_{ijkl} is the response, percent correct. Therefore, ideally, we would have $N = 3 \cdot 4 \cdot 11 \cdot 15 = 1980$ observations. Unfortunately, two rats became ill during the experiment and some equipment difficulties resulted in a number of missing session responses. Thus these data points are excluded from the analysis (hence $N = 1710$).

In terms of model (1), using vector notation and adding distributional assumptions for testing various components, we can rewrite (4) as

$$\mathbf{Y} \mid \mathbf{s} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{s}, \mathbf{R}) \tag{5}$$

where $\boldsymbol{\beta} = (\mu, \tau_1, \ldots, \tau_3, d_1, \ldots, d_4)^T$ is the vector of fixed effects, $\mathbf{s}$ is the 11×1 vector of random effects, $\mathbf{X}$ is an $N \times 8$ design matrix, $\mathbf{Z}$ is an $N \times 11$ design matrix, and $\mathbf{R}$ is an $N \times N$ covariance matrix. In order to estimate the effects we need to place restrictions on the vector $\boldsymbol{\beta}$. We will assume $\tau_3 = 0$ and $d_4 = 0$

as baseline constraints. Furthermore, we assume a normal hierarchy where $\mathbf{s}$ is a random variable such that

$$\mathbf{s} \sim N(0, \mathbf{G}) \tag{6}$$

as in (2) with $\mathbf{G}$ an 11×11 covariance matrix.

A number of remarks are necessary for justification of the model (5) and (6) in this context. First, let us consider the different factors in the model. We assume that the effect due to session, $\mathbf{s}$, is random. The motivation for such a supposition is that session is a time factor whereby all rats are observed on a particular session before the next session begins. Hence it restricts randomization in the experimental design. Additionally, measurements on a specific session may be more homogeneous than measurements between sessions and primary interest is in the treatment effect not session. Therefore we can think of the session factor as a block. Classically in a randomized complete block design, blocks are presumed to be random effects. As a side note, placing the effect of session further down in the hierarchy (stage two as defined by (2)) allows us to distinguish and study the correlation between observations on the same rat over time through the matrix $\mathbf{G}$. The assumed normal distribution for the random variable $\mathbf{s}$ is often made by investigators (Searle, Casella, and McCulloch (1991)) and guarantees a mathematically tractable and computationally feasible model to fit.

The effects due to delay, $\mathbf{d}$, are assumed to be fixed. In contrast to the session factor, we are interested only in the four particular levels of the intertrial delay employed in the experiment. Therefore, it is natural to consider the effects of these delays on rat performance as some unknown constant (i.e., fixed). Similarly, the effect due to treatment is presumed fixed since we are studying three specific levels of lead dosage applied in the experiment. The other factor we may wish to fit in our model is the box effect. The experimental apparatus is contained in one of ten boxes. The differences between boxes may explain some of the variation in the observed responses. However, upon placing a box variable in the model, we find it has no significant effect on rat performance. Hence, the effects due to box are not

included in the final analysis.

As a second remark, let us consider the covariance components of the model, **R** and **G** in (5) and (6). The standard ANOVA assumption of independent observations is likely to be violated due to the repeated measurements on the same rat across session and delay. The correlation over sessions is modeled by the covariance matrix **G**. This matrix, though, is 12×12 in dimension and hence may contain $\frac{12 \cdot 13}{2} = 78$ parameters if unrestricted. In order to make the computations for fitting the model feasible we need to place constraints on **G**. We thus assume an AR(1) structure where

$$G = \sigma^2 \cdot \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{11} \\ \rho & 1 & \rho & \dots & \rho^{10} \\ \vdots & & & & \vdots \\ \rho^{11} & \rho^{10} & \rho^9 & \dots & 1 \end{pmatrix}. \tag{7}$$

Here σ^2 is a positive variance component of **G** and $\rho \in (0, 1)$ is an unknown correlation term. This supposition reduces the number of unknown parameters in **G** from 78 to two. Intuitively, the AR(1) covariance structure models observations further apart in time as less correlated than observations closer together in time. Therefore, the AR(1) presumption is not only computationally desirable but also realistic.

The correlation due to the intertrial delay is modeled by the matrix **R**. This matrix has an interesting structure as a result of the following constraints. First, we assume that observations between rats are independent. Second, at the primary level of the hierarchy (5) we condition on realizations of the random variable session, s. As a consequence, since we have essentially made a random selection of the effects due to session and fixed them at this stage, the observations across time are independent. Another way of thinking about this independence is that the correlation due to session is accounted for in the second stage of the hierarchy (6). Therefore, the linear model in (5) needs to fit only the correlation across delay and not sessions. Third, we suppose the correlation between observations within a rat on a particular session is the same for each rat by session combination. In a sense, then,

we are considering the effect of delay on performance homogeneous across rats. $\mathbf{R}$ therefore models the correlation due to the repeated measurements over delay within rat by session. Subsequently, $\mathbf{R}$ is a block diagonal matrix with 495 equal blocks (corresponding to the $45 \cdot 11$ combinations of rat and session) of dimensions 4×4 (four delay levels):

$$R = \begin{pmatrix} \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22}^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33}^2 & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44}^2 \end{bmatrix} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22}^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33}^2 & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44}^2 \end{bmatrix} \end{pmatrix}. \tag{8}$$

Since we are presuming each block is identical, we do not need to restrict the number of parameters in each individual block. Hence, $\mathbf{R}$ contains $\frac{4 \cdot 5}{2} = 10$ unknown parameters for estimation.

The model (5) and (6) can be fit using Proc Mixed in Statistical Analysis System (SAS technical report, chapter 16). An annotated program for the implementation of the mixed model appears in the appendix. In this application, Proc Mixed estimates the covariance parameters using REML as discussed in Section 3. Empirical Bayes methodologies are utilized for estimates of the fixed effects. More specifically, Proc Mixed solves the mixed model equations given by Henderson (1984), with the REML estimates, $\hat{\mathbf{G}}$ and $\hat{\mathbf{R}}$, of $\mathbf{G}$ and $\mathbf{R}$ plugged in to calculate estimates for $\boldsymbol{\beta}$ and the realizations of $\mathbf{s}$. Furthermore, through these mixed model equations along with the observed Fisher information matrix, Proc Mixed computes approximate standard errors for estimates of the fixed effects, realized random effects, and covariance parameters and hence can perform various hypothesis tests of these parameters (see SAS technical report, chapter 16, for more details).

Table 1: F-tests of the fixed effects.

Source	df	F value	p-value
Treatment	2	8.69	0.0002
Delay	3	180.46	< 0.0001
Delay*Treatment	6	1.13	0.3400

Table 2: Tests of the planned contrasts.

Contrast	Estimate	Standard Error	T value	p-value
Trt. 0 vs. 75-300	2.62	0.67	3.97	0.0001
Trt. 75 vs. 300	0.51	0.38	1.36	0.1700
Trt. 0-75 vs. 300	2.08	0.66	3.15	0.0017

The tests of the fixed effects is shown in Table 1. The degrees of freedom in the denominator of the F-tests is $N - 1 = 1709$. There is a significant effect of lead treatment ($p = 0.0002$) and intertrial delay ($p < 0.0001$) on rat performances. In Table 2 we compute pairwise comparisons and find the control group (zero lead) differs significantly from combinations of the two lead groups ($p = 0.0001$). Likewise, the high lead group differs significantly from a combination of the low lead and control groups ($p = 0.0017$). However, the two lead groups do not differ significantly ($p = 0.17$). As for the effect of session on rat performance, Table 3 displays the conditional expected values of the session effects given the responses (estimate column) along with approximate t- tests (1709 degrees of freedom). It appears the rats are performing better on the memory task as the experimental days pass during the first "week" of the study. On the later sessions (8-11), rat performance seems to level off. This trend may indicate some kind of learning effect in that the rats are playing the game better as the investigation progresses. Be that as it may, these effects of session on rat memory are not significantly different than zero ($p > 0.59$ for each effect in Table 3).

Proc Mixed also presents REML estimates of the covariance parameters from **R**

Table 3: Solutions for the random effect, session.

Parameter	Estimate	Standard Error	T value	p-value
Session 1	-6.03	11.37	-0.53	0.60
Session 2	-2.03	11.39	-0.18	0.86
Session 3	0.72	11.40	0.06	0.95
Session 4	1.82	11.40	0.16	0.87
Session 5	2.52	11.40	0.22	0.83
Session 6	4.86	11.40	0.43	0.67
Session 7	5.64	11.40	0.49	0.62
Session 8	5.63	11.40	0.49	0.62
Session 9	5.09	11.40	0.45	0.66
Session 10	6.06	11.39	0.53	0.60
Session 11	5.54	11.37	0.49	0.63

and $\mathbf{G}$ in (5) and (6). Recall we assume $\mathbf{G}$ has an AR(1) structure as in (7) and $\mathbf{R}$ is block diagonal as in (8). From Table 4 we can see $\rho = 0.98$ is significantly different than zero ($p < 0.0001$) and $\sigma^2 = 139.34$ is a highly variable estimate of the variance of the session effect. The parameter estimates for the 4×4 blocks of $\mathbf{R}$ are also listed in Table 4, represented by $\text{UN}(i,j)$, $i,j \in \{1,\ldots,4\}$. Notice that the off-diagonal elements, $\text{UN}(i,j)$, $i \neq j$, are similar. This observation may lead us to conclude that an equicorrelation assumption is not unreasonable.

The estimates and inferences presented above can not be trusted without a legitimate model. Proc Mixed computes a number of statistics as model fitting information. For example, we can obtain a deviance or log likelihood value from our model. This deviance can be used for comparison of the model to various submodels with more constraints. As an illustration, SAS tests the present model against the null model of independent error terms ($\mathbf{R} = \sigma^2\mathbf{I}$). Using a likelihood ratio χ^2 test statistic we find the current model is preferred to the ordinary least squares type null model ($p < 0.0001$ on eleven degrees of freedom (12 parameters from Table 4 minus one for the independence model)). Furthermore, we examined the standard plots of residuals vs. predicted values and residuals vs. normal scores to check

Table 4: REML estimates of the covariance parameters from **R** and **G**.

G cov. param.	Estimate	Standard Error	Z value	p-value
Diagonal	139.34	275.66	0.51	0.6100
AR(1)	0.98	0.033	29.97	< 0.0001
R cov. param.				
UN(1,1)	102.45	6.98	14.67	< 0.0001
UN(2,1)	10.98	5.48	2.01	0.0450
UN(2,2)	125.01	8.58	14.56	< 0.0001
UN(3,1)	12.64	5.99	2.11	0.0350
UN(3,2)	18.70	6.69	2.79	0.0052
UN(3,3)	150.84	10.22	14.76	< 0.0001
UN(4,1)	13.87	5.87	2.36	0.0180
UN(4,2)	4.68	6.53	0.72	0.4700
UN(4,3)	13.46	6.95	1.94	0.0530
UN(4,4)	133.14	9.35	14.24	< 0.0001

for potential violations in the normality and (rat) independence assumptions. The graphs did not indicate any deviations from these suppositions. However, it is not clear whether the residual plots have the same interpretations under this complex model as compared to a more standard general linear model (see Section 5 for more remarks on model diagnostics).

In summary, the model (5) and (6) indicates there is a significant effect of low lead dosages on rat performance. In particular, the two lead groups perform significantly worse on the memory task than the control group, though one lead group does not significantly outperform the other. Furthermore, our assumption that the rats learned the rules of the game prior to the experiment needs further study as to its validity.

5 Discussion

The mixed model (5) and (6) presented in the previous section provides a way of accounting for the correlation structure inherent in a repeated measures design thus

enabling us to study the effect of lead on memory in rats. There are a number of issues surrounding the model that need to be addressed and studied further. First of all, the model considers the effects of session on performance as random. However, we are not obtaining a random sample from a population of sessions over which to run the experiment. In fact, we have a group of consecutive days (or at least eleven days clumped within a two week span) chosen for convenience of data collection. Though we may wish to infer to the population of all session effects, the manner in which the observations are gathered over time may prompt us to assume fixed session effects. Furthermore, in Section 2 we show that the model for analyzing data from a split-unit design is a special case of the mixed model (1) and (2) with the session factor as fixed and the rat effect as random. Hence, fitting a mixed model with fixed session effects may not only be interesting for comparison with the model developed in Section 4, but also as a means for validating some of the assumptions made by alternative models such as that from a split-unit design.

The difficulty with representing the session factor as fixed is that the $\mathbf{R}$ matrix will model the correlation due to observations within rats on a particular session over different intertrial delays *and* observations within rats over different sessions. Without restrictions on the correlation structure, the computation of estimators for the resulting large number of covariance parameters may be unmanageable. Therefore, additional consideration is necessary for finding realistic constraints to place on $\mathbf{R}$ in order to take advantage of the built-in covariance structures in Proc Mixed and make the model computationally feasible.

A second question of interest concerning the mixed model in Section 4 (and in general actually) is that of model validation. How can we determine whether the presumed normal hierarchy is correct or the covariance structure assumed for $\mathbf{R}$ and $\mathbf{G}$ are reasonable? The complexity of the model may make standard tests of fit hard to implement. For example, SAS computes the maximum value for the log likelihood of our model. This quantity can then be compared to similar quantities from a larger model through a χ^2 type statistic. Alternatively we can determine if it

is satisfactory to fit a reduced model over the model under consideration. The variety of covariance structures available in Proc Mixed will allow us, pending computing limitations, to fit such less restricted models or further constrained models. In our modeling situation, for instance, we can fit an equicorrelated structure over delay (use TYPE=CS in the REPEATED statement of the SAS code presented in the Appendix) and test whether the model fit in Section 4 is significantly different than the constrained model. However, competing models may have completely different covariance structures that do not lend themselves to such a testing routine. In particular, we can not compare the split-unit model, say, with the model fit in Section 4 using this method. Evaluation based on realism and expertise in the area may be our best recourse.

As another example, we can compute residual plots and similar diagnostics to check validity of our distribution assumptions. The basic concept behind the examination of residuals is to determine if they exhibit tendencies contrary to model assumptions. However, unlike the standard ANOVA where error terms are presumed independent and normal, it is not clear how deviations from model assumptions will affect the behavior of the residuals under our complicated error structure. Further research is necessary to establish whether residual diagnostics can easily be conducted for mixed model validation.

Lastly, in our analysis we summarize data over trials run on a particular session in terms of a percent correct. This summary may result in the loss of relevant information. Ideally, we would like to consider each individual trial as an observation. Since a trial consists of a rat making either a correct or incorrect nose poke, a logistic model with a binary 0-1 response would be appropriate. This approach provides an alternative to the normal hierarchy assumed in our model.

The model (5) and (6) is not necessarily the best one for the data at hand. Nevertheless, it does allow us to study the treatment of interest (lead) while reasonably accounting for other variables (i.e. intertrial delay and session) effecting the response (rat performance). In comparison to the split-unit, "curve-fitting," many ANOVAs,

and averaging-ANOVA techniques discussed in Section 2, the mixed model approach provides a more general means for handling the correlation between observations over the repeated measures. We can then obtain a more thorough analysis of the treatment effects, time trends, and interactions of treatment with the other factors of interest. Moreover, the mixed model framework is flexible enough to allow for different modeling strategies. For example, the $\mathbf{R}$ and $\mathbf{G}$ matrices can be chosen to fit any type of correlation structure in the data with which a researcher might be presented.

On the computing end, the mixed model is more complex than the procedures discussed in Section 2. However, available software such as Proc Mixed in SAS version 6.09 makes the model fitting task feasible. The data analysis in Section 4 takes 14.35 minutes on a SPARC Sun station and is not too difficult to program and implement in SAS (see Appendix). Additionally, the mixed model is a general linear model (GLM) with correlated error structure. Consequently, the output generated by SAS is similar to that of other GLM routines (e.g. ANOVA) and is hence not hard to interpret or study.

As a final note, the GLM framework of the mixed model broadens its' applicability beyond that of just analyses of repeated measures type designs. In fact, any experiment consisting of random and fixed factors with correlated observations can be studied through the mixed model. For the above reasons, the gain in flexibility and realism with the mixed model seems to outweigh the increase in complexity and computational difficulties. Therefore, the mixed model appears to be a good inferential tool for analyzing the effect of lead on memory in rats as well as many other similarly designed experiments.

Acknowledgments

Richard A. Levine is supported by the Office of Naval Research Graduate Fel-

lowship and NIH Training Grant No. ES07261. We would like to thank Professors George Casella and David Ruppert for many helpful discussions and advise through the course of this research and comments on drafts of this manuscript. We would also like to thank Susan Alber and Professor Barbara Strupp for providing the data on low lead exposure in rats and for encouraging us to understand the statistics from a practical viewpoint.

6 Appendix

```
/*  This is the SAS version 6.09 code for implementing the analysis described
/*  in Section 4.  This program was run on a SunOS UNIX 10/51 */

/*  Initialize for output and input */

options ls=70;
libname here '/user2/ral4';
title  'Assessing the Effect of Low-level Lead Exposure';

/*  Input the data and sort it */

data data1;
infile '/user2/ral4/summ7a.dat';

INPUT  treatment session delay percentcorrect n rat $;
PROC SORT data=data1 out=here.summ7a ;
BY rat session delay;

/*  Proc Mixed call
/*  CONVF, CONVH,  and CONVG set the convergence criteria and we ask
/*  for REML estimation of the unknown parameters.  */

PROC MIXED METHOD=REML CONVF=1E-5 CONVG=1E-5 CONVH=1E-5
DATA=here.summ7a;

CLASS  session delay treatment rat;    /*  declare the discrete variables  */

/*  The mixed model: treatment, delay, and the delay by treatment interaction
/*  are fixed effects in the model statement.
/*  Session is assumed to be a random effect so is placed in the RANDOM
/*  statement.  Conditional on realizations of the session effect, repeated
/*  measures are still taken over delay within rat*session.
/*  The REPEATED statement models this situation.
```

```
/*  Explanation of commands after the backslash, /, on each line:
/*  S and CL asks Proc Mixed to print solutions for the estimate effects
/*  and 95% confidence intervals.
/*  G and R request for print outs of the
/*  G and R matrices respectively.
/*  TYPE allows for specification of the correlation structure
/*  (AR(1) for G and unrestricted (UN) for R).
/*  SUBJECT forces a block diagonal matrix with identical blocks for each
/*  subject (i.e., identical blocks here for each rat by session). */

MODEL percentcorrect = treatment delay treatment*delay / S CL;
RANDOM session / TYPE=AR(1) S G CL;
REPEATED delay / TYPE=UN SUBJECT=rat*session R ;

/* Compute least squares means and confidence intervals of
/* treatment, delay, and the delay by treatment interaction  */

LSMEANS treatment delay treatment*delay / CL;

/*  Compute treatment contrasts  */

/* control vs. lead */
ESTIMATE 'TRT 0 VS 75-300' treatment 1 -0.5 -0.5;

/* compare two lead groups */
ESTIMATE 'TRT 75 VS 300' treatment 0  0.5 -0.5;

/* zero and low lead  vs. high lead */
ESTIMATE 'TRT 0-75 VS 300' treatment 0.5 0.5 -1;     run;
```

References

[1] Angell, N. F. and Weiss, B. (1982). Operant Behavior of Rats Exposed to Lead Before or After Weaning. *Toxicology and Applied Pharmacology* 63, 62-71.

[2] Barrett, J. and Livesey, P. J. (1985). Low Level Lead Effects on Activity Under Varying Stress Conditions in the Developing Rat. *Pharmacology Biochemistry and Behavior* 22, 107-118.

[3] Brown, D. R. (1975). Neonatal Lead Exposure in the Rat: Decreased Learning as a Function of Age and Blood Lead Concentrations. *Toxicology and Applied Pharmacology* 32, 628-637.

[4] Brown, S., Dragann, N. and Vogel, W. H. (1971). Effects of Lead Acetate on Learning and Memory in Rats. *Archives of Environmental Health* 22, 370-372.

[5] Bushnell, P. J. and Bowman, R. E. (1979). Persistence of Impaired Reversal Learning in Young Monkeys Exposed to Low Levels of Dietary Lead. *Journal of Toxicology and Environmental Health* 5, 1015-1023.

[6] Bushnell, P. J. and Bowman, R. E. (1979). Reversal Learning Deficit in Young Monkeys Exposed to Lead. *Pharmacology Biochemistry and Behavior* 10, 733-742.

[7] Carson, T. L., Van Gelder, G. A. , Karas, G. C. and Buck, W. B. (1974). Slowed Learning in Lambs Prenatally Exposed to Lead. *Archives of Environmental Health* 29, 154-156.

[8] Cochran, W. G. and Cox, G. M. (1957). *Experimental Designs.* Second Edition. John Wiley & Sons, New York.

[9] Cory-Slechta, D. H., Bissen, S. T. ,Young, A. M. and Thompson, T. (1981). Chronic Postweaning Lead Exposure and Response Duration Performance. *Toxicology and Applied Pharmacology* 60, 78-84.

[10] Cory-Slechta and Thompson, T. (1979). Behavioral Toxicity of Chronic Postweaning Lead Exposure in the Rat. *Toxicology and Applied Pharmacology* 47, 151-159.

[11] Cory-Slechta, Weiss, B. ,and Cox, C. (1983). Delayed Behavioral Toxicity of Lead with Increasing Exposure Concentration. *Toxicology and Applied Pharmacology* 71, 342-352.

[12] Cory-Slechta, Weiss, B. , and Cox, C. (1985). Performance and Exposure Indices of Rats Exposed to Low Concentrations of Lead. *Toxicology and Applied Pharmacology* 78, 291-299.

[13] Davis, J. M., Otto, D. A., Weil, D. E., and Grant, L. D. (1990). The Comparative Developmental Neurotoxicity of Lead in Humans and Animals. *Neurotoxicology and Teratology* 12, 215-229.

[14] Davis, C. S. and Wei, L. J. (1988). Nonparametric Methods for Analyzing Incomplete Nondecreasing Repeated Measurements. *Biometrics* 44, 1005-1018.

[15] Dietrich, K. N., Berger, O. G., and Succop, P. A. (1993). Lead Exposure and the Motor Developmental Status of Urban Six-Year-Old Children in the Cincinnati Prospective Study. *Pediatrics* 91 (2), 301- 307.

[16] Driscoll, J. W. and Stegner, S. E. (1976). Behavioral Effects of Chronic Lead Ingestion on Laboratory Rats. *Pharmacology Biochemistry and Behavior* 4, 411-417.

[17] Dwyer, J. H., Feinleib, M., Lippert, P., and Hoffmeister, H. (1991). *Statistical Models for Longitudinal Studies of Health.* Oxford University Press, New York.

[18] Fitzmaurice, G. M., Laird, N. M., and Rotnitzky, A. G. (1993). Regression Models for Discrete Longitudinal Responses. *Statistical Science* 8 (3), 284-309.

[19] Freedman, D. A. and Zeisel, H. (1988). From Mouse-to-Man: The Quantitative Assessment of Cancer Risks. *Statistical Science* 3 (1), 3-56.

[20] Gilbert, S. G. and Rice, D. C. (1987). Low-Level Lifetime Lead Exposure Produce Behavioral Toxicity (Spatial Discrimination Reversal) in Adult Monkeys. *Toxicology and Applied Pharmacology* 91, 484-490.

[21] Grant, L. D., Kimmel, C. A. ,West, G. L. , Martinez-Kargas, C. M. and Howard, J. L. (1980). Chronic Low-Level Toxicity in the Rat II. Effects on Postnatal Physical and Behavioral Development. *Toxicology and Applied Pharmacology* 56, 42-58.

[22] Greenhouse, S. W. and Geisser, S. On Methods in the Analysis of Profile Data. *Psychometrika* 24, 95-112.

[23] Grizzle, J. E. and Allen, D. M. (1969). Analysis of Growth and Dose Response Curves. *Biometrics* 25 (2), 357-381.

[24] Hastings, L., Cooper, G. , Bornschein, R. and Michaelson, I. A. (1977). Behavioral Effects of Low Level Neonatal Lead Exposure. *Pharmacology Biochemistry and Behavior* 7, 37-42.

[25] Henderson, C. R. (1984). *Applications of Linear Models in Animal Breeding.* University of Guelph.

[26] Hunter, J., Urbanowicz, M. A., Yule, W., and Lansdown, R. (1985). Automated Testing of Reaction Time and Its Association with Lead in Children. *International Archives of Occupational and Environmental Health* 57 (1), 27-34.

[27] Jones, R. H. (1993). *Longitudinal Data with Serial Correlation: A State-Space Approach.* Chapman and Hall.

[28] Koch, G. G., Elashoff, J. D., Amara, I. A. (1988). Repeated Measurements- Design and Analysis. *Encyclopedia of Statistical Sciences* 8, 46-73.

[29] Koziol, J. A., Maxwell, D. A., Fukishima, M., Colmerauer, M. E. M., and Pilich, Y. H. (1981). A Distribution-free Test for Tumor-Growth Curve Analyses with Application to an Animal Tumor Immunotherapy Experiment *Biometrics* 37, 383-390.

[30] Laird, N. M. and Ware, J. H. (1982). Random-Effects Model for Longitudinal Data. *Biometrics* 38, 963-974.

[31] Lehmann, E. L. (1983). *Theory of Point Estimation.* Wadsworth, California.

[32] Lentner, M. and Bishop, T. (1993). *Experimental Design and Analysis*, 2nd Edition. Valley Book Company, Virginia.

[33] Levin, E. D. and Bowman, R. E. (1986). Long-Term Lead Effects on the Hamilton Search Task and Delayed Alternation in Monkeys. *Neurotoxicology and Teratology* 8, 219-224.

[34] Levin, E. D. and Bowman, R. E. (1986). Scopolamine Effects on Hamilton Search Task Performance in Monkeys. *Pharmacology Biochemistry and Behavior* 24, 819-821.

[35] Levin, E. D. and Bowman, R. E. (1988). Long-Term Lead Effects of Postnatal Lead Exposure on Delayed Spatial Alternation in Monkeys. *Neurotoxicology and Teratology* 10, 505-510.

[36] Lilienthal, H., Winneke, G. , Brockhaus, A. and Molik, B. (1986). Pre- and Postnatal Lead-Exposure in Monkeys: Effects on Activity and Learning Set Formation. *Neurotoxicology and Teratology* 8, 265-272.

[37] Lindstrom, M. J. and Bates, D. M. (1988). Newton-Raphson and EM Algorithms for Linear Mixed-effects Model for Repeated Measures Data *Journal of American Statistical Association* 83, 1014-1022.

[38] Massaro, T. F., Miller, G. D. and Massaro, E. J. (1986). Low-Level Lead Exposure Affects Latent Learning in Rat. *Neurotoxicology and Teratology* 8, 109-113.

[39] Mead, R. (1988). *The Design of Experiments.* Cambridge University Press.

[40] Müller, H. G. (1988). *Nonparametric Regression Analysis of Longitudinal Data.* Springer- Verlag.

[41] Needleman, H. L. (1994). Preventing Childhood Lead Poisoning. *Preventive Medicine* 23 (5), 634-637.

[42] Neter, J., Wasserman, W., and Kutner, M. H. (1990). *Applied Linear Statistical Models.* Third edition, Irwin, Inc.

[43] Overmann, S. R. (1977). Behavioral Effects of Asymptomatic Lead Exposure during Neonatal Development in Rats. *Toxicology and Applied Pharmacology* 41, 459-471.

[44] Potthoff, R. F. and Roy, S. N. (1964). A Generalized Multivariate Analysis of Variance Model Useful Especially for Growth Curve Problems. *Biometrika* 51, 313-326.

[45] Rao, C. R. (1965). The Theory of Least Squares when the Parameters Are Stochastic and Its Application to the Analysis of Growth Curves. *Biometrika* 52, 447-458.

[46] Reiter, L. W., Anderson, G. E. , Laskey, J. W. and Cahill, D. F. (1975). Developmental and Behavioral Changes in the Rat during Chronic Exposure to Lead. *Environmental Health Perspectives* 12, 119- 123.

[47] Rice, D.C. (1984). Behavioral Deficit (Delayed Matching to Sample) in Monkeys Exposed from Birth to Low Levels of Lead. *Toxicology and Applied Pharmacology* 75, 337-345.

[48] Rice, D.C. (1984). Chronic Low-Lead Exposure from Birth Produces Deficits in Discrimination Reversal in Monkeys. *Toxicology and Applied Pharmacology* 77, 201-210.

[49] Rice, D.C. (1990). Lead-Induced Behavioral Impairment a on a Spatial Discrimination Reversal Task in Monkeys Exposed during Different Periods of Development. *Toxicology and Applied Pharmacology* 106, 327-333.

[50] Rice, D.C. (1992). Behavioral Effects of Lead in Monkeys Tested during Infancy and Adulthood. *Neurotoxicology and Teratology* 14, 235-245.

[51] Rice, D.C. and Gilbert, S. G. (1985). Low Lead Exposure from Birth Produces Behavioral Toxicity (DRL) in Monkeys. *Toxicology and Applied Pharmacology* 80, 421-426.

[52] Rice, D.C. and Gilbert, S. G. (1990). Sensitive Periods for Lead-Induced Behavioral Impairment (Nonspatial Discrimination Reversal) in Monkeys. *Toxicology and Applied Pharmacology* 102, 101-109.

[53] Rice, D.C. and Gilbert, S. G. (1990). Lack of Sensitive Periods for Lead-Induced Behavioral Impairment on a Spatial Delayed Alternation Task in Monkeys. *Toxicology and Applied Pharmacology* 103, 364-373.

[54] Rice, D.C. and Karpinski, K. F. (1988). Lifetime Low-Level Lead Exposure Produces Deficits in Delayed Alternation in Adult Monkeys. *Neurotoxicology and Teratology* 10, 207-214.

[55] Sandland, R. L. and McGilchrist, C. A. (1979). Stochastic Growth Curve Analysis. *Biometrics* 35, 255-271.

[56] SAS Institute Inc. (1993). SAS/STAT Software: Changes and Enhancements, Release 6.07. Technical report no. P-229. Cary, NC: SAS Institute Inc.

[57] Searle, S. R., Casella, G. and McCulluoch, C. E. (1992). *Variance Components*, John Wiley & Sons, New York.

[58] Snedecor, G. W. and Cochran, W. G. (1989). *Statistical Methods.* Eighth Edition. Iowa State University Press, Iowa.

[59] Snowdon, C. T. (1973). Learning Deficits in Lead-Injected Rats. *Pharmacology Biochemistry and Behavior* 1, 599-603.

[60] Stanek, E. J. and Diehl, S. R. (1988). Growth Curve Models of Repeated Binary Response. *Biometrics* 44, 973-983.

[61] Tanner, M. A. (1993). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, 2nd edition. Springer-Verlag, New York.

[62] Thompson, W. A., Jr. (1962). The Problem of Negative Estimates of Variance Components. *Annals of Mathematical Statistics* 33, 273-289.

[63] Vonesh, E. F. and Carter, R. L. (1987). Efficient Inference for Random-Coefficient Growth Curve Models with Unbalanced Data. *Biometrics* 43, 617-628.

[64] Ware, J. H. (1985). Linear Models for the Analysis of Longitudinal Studies. *The American Statistician* 39 (2), 95-101.

[65] Waternaux, C., Laird, N. M., and Ware, J. H. (1989). Methods for Analysis of Longitudinal Data: Blood- Lead Concentrations and Cognitive Development. *Journal of the American Statistical Association* 84, 33-41.

[66] Wei, L. J. and Johnson, J. E. (1985). Combining Dependent Tests with Incomplete Repeated Measurements. *Biometrika* 72, 359-364.

[67] White, R. F., Diamond, R., Proctor, S., Morey, C., and Hu, H. (1993). Residual Cognitive Deficits 50 Years After Lead Poisoning During Childhood. *British Journal of Industrial Medicine* 50 (7), 613- 622.

[68] Wishart, J. (1938). Growth Rate Determinations in Nutrition Studies with the Bacon Pig, and Their Analysis. *Biometrika* 30, 16-28.

[69] Zeger, S. L., Liang, K. Y., and Albert, P. S. (1988). Models for Longitudinal Data: A Generalized Estimating Equation Approach. *Biometrics* 44, 1049-1060.